10 - Multivariate Analysis of Variance

Junvie Pailden

SIUE, F2017, Stat 589

September 18, 2017

Comparing Several Mult Pop'n Means (Multivariate ANOVA)

Random samples, collected from each of g populations,

Population 1:	$\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}$
Population 2:	$\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2}$
$\vdots$	$\vdots$
Population g:	$\mathbf{X}_{g1}, \mathbf{X}_{g2}, \ldots, \mathbf{X}_{gn_g}$

- MANOVA is used first to investigate whether the population mean vectors are the same and, if not, which mean components differ significantly.

Review of Univariate ANOVA

- $X_{\ell 1}, X_{\ell 2}, \ldots, X_{\ell n_l}$ is a random sample from an $N(\mu_\ell, \sigma^2)$ population, $\ell = 1, 2, \ldots, g$
- random samples are independent
- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_q$
- $\mu_\ell = \mu + (\mu_\ell - \mu) = \mu + \tau_\ell$, where $\tau_\ell = \mu_\ell - \mu$.
- The null hypothesis becomes $H_0 : \tau_1 = \tau_2 = \cdots = \tau_g = 0$
- The response $X_{\ell j} \sim N(\mu + \tau_j, \sigma^2)$, can be written as

$$X_{\ell j} = \mu + \tau_\ell + e_{\ell j}$$

  where the $e_{\ell j}$ are independent $N(0, \sigma^2)$ random variables.
- To define uniquely the model parameters and their least squares estimates, we impose the constraint

$$\sum_{\ell=1}^{g} n_\ell \tau_\ell = 0.$$

The analysis of variance is based upon an analogous decomposition of the observations

$$
\begin{aligned}
x_{\ell j} &= \bar{x} + (\bar{x}_\ell - \bar{x}) + (x_{\ell j} - \bar{x}_\ell) \\
&= \bar{x} + \hat{\tau}_\ell + \hat{e}_{\ell j} \\
\text{(obs)} &= \text{(overall sample mean)} + \\
&\qquad \text{(estimated treatment effect)} + \text{(residual)}
\end{aligned}
$$

Note that, for all $\ell = 1, 2, \ldots, g$,

$$
\sum_{\ell=1}^{n_\ell} (x_{\ell j} - \bar{x})^2 = n_\ell (\bar{x}_\ell - \bar{x})^2 + \sum_{j=1}^{n_\ell} (x_{\ell j} - \bar{x}_\ell)^2,
$$

since $\sum_{j=1}^{n_\ell} (x_{\ell j} - \bar{x}_\ell) = 0$.

Summing both sides over ℓ we get

$$\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(x_{\ell j}-\bar{x})^2 = \sum_{\ell=1}^{g} n_\ell(\bar{x}_\ell-\bar{x})^2 + \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(x_{\ell j}-\bar{x}_\ell)^2$$

$$\begin{pmatrix} SS_{cor} \\ \text{total} \\ \text{(corrected) SS} \end{pmatrix} = \begin{pmatrix} SS_{tr} \\ \text{between} \\ \text{(samples) SS} \end{pmatrix} + \begin{pmatrix} SS_{res} \\ \text{within} \\ \text{(samples) SS} \end{pmatrix}$$

OR

$$\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell} x_{\ell j}^2 = (n_1+n_2+\cdots+n_g)\bar{x}^2 + \sum_{\ell=1}^{g} n_\ell(\bar{x}_\ell-\bar{x})^2 + \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(x_{\ell j}-\bar{x}_\ell)^2$$

$$(SS_{obs}) = (SS_{mean}) + (SS_{tr}) + (SS_{res})$$

ANOVA Table

Source of variation	Sum of Squares(SS)	Degrees of freedom($d.f.$)
Treatments	$SS_{tr} = \sum_{\ell=1}^{g} n_\ell (\bar{x}_\ell - \bar{x})^2$	$g - 1$
Residual (error)	$SS_{res} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (x_{\ell j} - \bar{x}_\ell)^2$	$\sum_{\ell=1}^{g} n_\ell - g$
Total (corrected for the mean)	$SS_{cor} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (x_{\ell j} - \bar{x})^2$	$\sum_{\ell=1}^{g} n_\ell - 1$

ANOVA Test for Comparing Univariate Means

- The usual F-test rejects $H_0 : \tau_1 = \tau_2 = \cdots = \tau_g = 0$ at level α if

$$F = \frac{SS_{tr}/(g-1)}{SS_{res}/\left(\sum_{\ell=1}^{g} n_\ell - g\right)} > F_{g-1, \sum n_\ell - g}(\alpha)$$

  where $F_{g-1, \sum n_\ell - g}(\alpha)$ is the upper (100α)th percentile of the F-distribution with $g-1$ and $\sum n_\ell - g$ degrees of freedom.
- This is equivalent to rejecting H_0 for large values of SS_{tr}/SS_{res} or for large values of $1 + SS_{tr}/SS_{res}$.
- The multivariate generalization rejects H_0 for small values of the reciprocal

$$\frac{1}{1 + SS_{tr}/SS_{res}} = \frac{SS_{res}}{SS_{res} + SS_{tr}}$$

Multivariate Analysis of Variance (MANOVA)

- MANOVA Model for Comparing g Population Mean Vectors

$$\mathbf{X}_{\ell j} = \mu + \tau_\ell + \mathbf{e}_{\ell j}, \;\; j = 1, 2, \ldots, n_\ell \text{ and } \ell = 1, 2, \ldots, g$$

  where the $\mathbf{e}_{\ell j}$ are independent $N_p(0, \Sigma)$ variables.
- The parameter vector μ is an overall mean (level), and τ_ℓ represents the ℓth treatment effect with $\sum_{\ell=1}^{g} n_\ell \tau_\ell = 0$.

A vector of observations may be decomposed

$$\mathbf{x}_{\ell j} = \bar{\mathbf{x}} + (\bar{\mathbf{x}}_\ell - \bar{\mathbf{x}}) + (\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_\ell)$$

$$\text{(observation)} = \begin{pmatrix} \text{overall sample} \\ \text{mean } \ \hat{\mu} \end{pmatrix} + \begin{pmatrix} \text{estimated} \\ \text{treatment} \\ \text{effect } \hat{\tau}_\ell \end{pmatrix} + \begin{pmatrix} \text{residual} \\ \hat{\mathbf{e}}_{\ell j} \end{pmatrix}$$

Similarly, we have

$$\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(\mathbf{x}_{\ell j}-\bar{\mathbf{x}})(\mathbf{x}_{\ell j}-\bar{\mathbf{x}})' = \sum_{\ell=1}^{g} n_\ell(\bar{\mathbf{x}}_\ell-\bar{\mathbf{x}})(\bar{\mathbf{x}}_\ell-\bar{\mathbf{x}})' + \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(\mathbf{x}_{\ell j}-\bar{\mathbf{x}}_\ell)($$

$$\begin{pmatrix} SS_{cor} \\ \text{total} \\ \text{(corrected) SS} \end{pmatrix} = \begin{pmatrix} SS_{tr} \\ \text{between} \\ \text{(samples) SS} \end{pmatrix} + \begin{pmatrix} SS_{res} \\ \text{within} \\ \text{(samples) SS} \end{pmatrix}$$

$$= \mathbf{B} + \mathbf{W}$$

The within sum of squares and cross products matrix can be expressed as

$$\begin{aligned}\mathbf{W} &= \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_\ell)(\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_\ell)' \\ &= (n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \cdots + (n_g - 1)\mathbf{S}_g\end{aligned}$$

where $\mathbf{S}_\ell$ is the sample covariance matrix for the ℓth sample.

MANOVA Table

The hypothesis of no treatment effects

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_g = 0$$

is tested by considering the relative sizes of the treatment and residual sums of squares and cross products.

Matrix of sum of squares and cross products(SSP)	Degrees of freedom$(d.f.)$
Treatments: $\mathbf{B} = \sum_{\ell=1}^{g} n_\ell (\bar{\mathbf{x}}_\ell - \bar{\mathbf{x}})(\bar{\mathbf{x}}_\ell - \bar{\mathbf{x}})'$	$g - 1$
Residual: $\mathbf{W} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_\ell)(\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_\ell)'$	$\sum_{\ell=1}^{g} n_\ell - g$
Total: $\mathbf{B} + \mathbf{W} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell j} - \bar{\mathbf{x}})(\mathbf{x}_{\ell j} - \bar{\mathbf{x}})'$	$\sum_{\ell=1}^{g} n_\ell - 1$

Wilks' Lamba Λ^*

- One test of

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_g = 0$$

  involves generalized variances. We reject H_0 if the ratio of generalized variances

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|} = \frac{\left|\sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_\ell)(\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_\ell)'\right|}{\left|\sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell j} - \bar{\mathbf{x}})(\mathbf{x}_{\ell j} - \bar{\mathbf{x}})'\right|}$$ is too small.

- The quantity Λ^* originally by Wilks corresponds to the equivalent form of the F-test of H_0: no treatment effects in the univariate case.

Wilks' Lamba Λ^* (cont)

- Wilks Λ^* can also be expressed as a function of the eigenvalues $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_s$ of $\mathbf{W}^{-1}\mathbf{B}$ as

$$\Lambda^* = \Pi_{i=1}^{s} \left(\frac{1}{1+\hat{\lambda}_i} \right), \text{ where } s = \min(p, g-1) = rank(B)$$

Distribution of Wilks' Lambda, Λ^*

No. of variables	No. of groups	Sampling distribution for multivariate normal data
$p = 1$	$g \geq 2$	$\left(\frac{\sum n_\ell - g}{g-1}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{g-1, \sum n_\ell - g}$
$p = 2$	$g \geq 2$	$\left(\frac{\sum n_\ell - g - 1}{g-1}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2(g-1), 2(\sum n_\ell - g - 1)}$
$p \geq 1$	$g = 2$	$\left(\frac{\sum n_\ell - p - 1}{p}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{p, \sum n_\ell - g - 1}$
$p \geq 1$	$g = 3$	$\left(\frac{\sum n_\ell - p - 2}{p}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p, 2(\sum n_\ell - p - 2)}$

Distribution of Wilks' Lambda, Λ^* (cont)

For other cases and large sample sizes, $\sum_\ell n_\ell = n$ large, we reject H_0 at significance level α if

$$-\left(-n - 1 - \frac{(p+g)}{2}\right) \ln \Lambda^* > \chi^2_{p(g-1)}(\alpha) \text{ or}$$

$$\Lambda^* < exp\left[-\left(-n - 1 - \frac{(p+g)}{2}\right)^{-1} \chi^2_{p(g-1)}(\alpha)\right]$$

Rootstock Data

The data contains four dependent variables as follows:

- six different rootstocks (Tree Number)
- trunk girth at four years (mm $\times$ 100)
- extension growth at four years (m)
- trunk girth at 15 years (mm $\times$ 100)
- weight of tree above ground at 15 years (lb $\times$ 1000)

```r
library(ACSWR) # data is in package ACSWR
data("rootstock")
colnames(rootstock) <- c("Tree.Num", "Girth.4y",
  "Growth.4y", "Girth.15y", "WgtAbvGrnd.15y")
```

```r
head(rootstock)
```

```
#   Tree.Num Girth.4y Growth.4y Girth.15y WgtAbvGrnd.15y
# 1        1      1.1       2.6       3.6           0.76
# 2        1      1.2       2.9       3.8           0.82
# 3        1      1.1       2.9       3.9           0.93
# 4        1      1.2       3.8       3.9           1.01
# 5        1      1.1       3.0       3.6           0.77
# 6        1      1.1       2.3       3.5           0.73
```

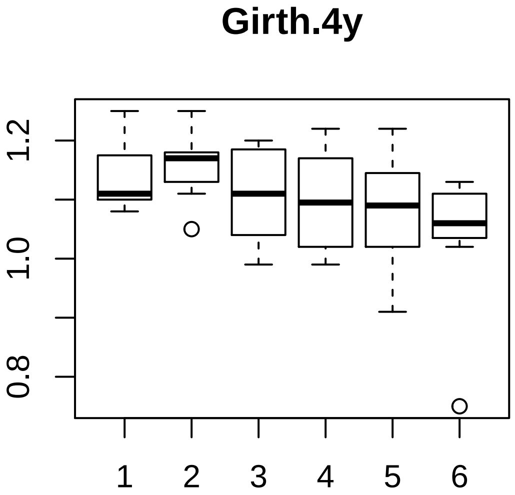
Setting up the data

The `manova()` function in R accepts formula interface `y ~ x`, where y is the matrix of dependent variables (measurement value) and x as the independent factor variable (population tree number).

```
dep.variable <- as.matrix(rootstock[, 2:5])
ind.variable <- as.factor(rootstock[, 1])
```

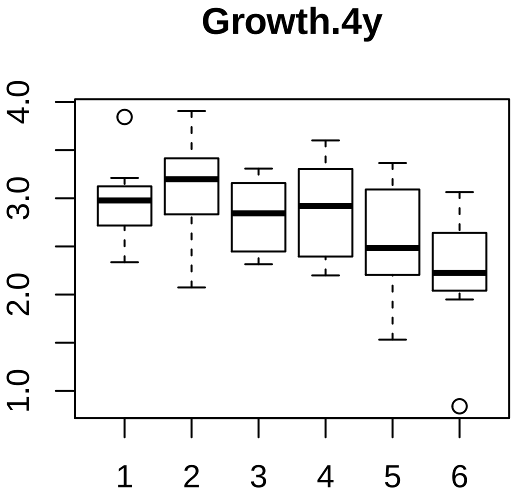
Side-by-Side Boxplots for Girth at 4 yrs

```
boxplot(Girth.4y ~ Tree.Num, data = rootstock,
                main = "Girth.4y")
```



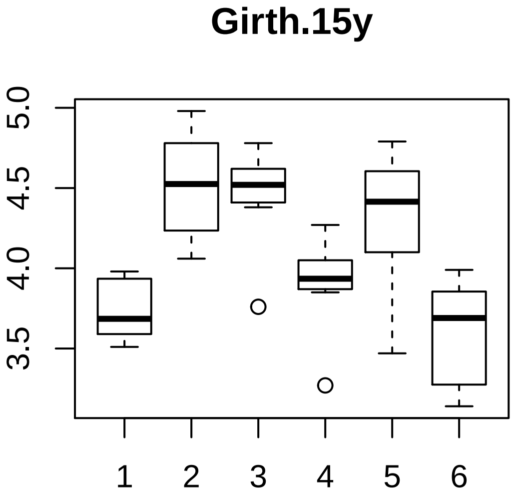
Side-by-Side Boxplots for Growth at 4 yrs

```
boxplot(Growth.4y ~ Tree.Num, data = rootstock,
            main = "Growth.4y")
```



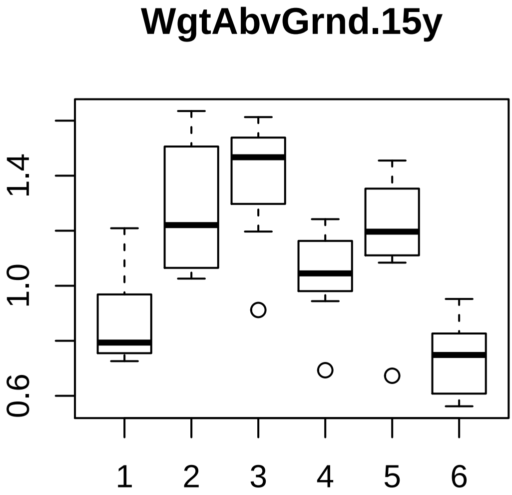
Side-by-Side Boxplots for Girth at 15 yrs

```
boxplot(Girth.15y ~ Tree.Num, data = rootstock,
            main = "Girth.15y")
```



Side-by-Side Boxplots for Weight Above Ground at 15 yrs

```
boxplot(WgtAbvGrnd.15y ~ Tree.Num, data = rootstock,
            main = "WgtAbvGrnd.15y")
```



MANOVA Test in R

y is the matrix of dependent variables (measurement value); x as the independent factor variable (population tree number).

```
rootstock.model <- manova( dep.variable ~ ind.variable )
summary(rootstock.model, test = "W") # argument test = W
```

```
#                Df Wilks approx F num Df den Df   Pr(>F)
# ind.variable    5 0.154     4.94     20    130 7.7e-09 ***
# Residuals      42
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

The MANOVA procedure gives a Wilks' test statistic of 0.154 and a p-value below 0.05, thus H_0 is rejected and it is concluded there are significant differences in the means measurements of the six different rootstocks.

Equivalent Test Statistics

In the context of random samples from several populations, the multivariate tests are based on the matrices

$$\mathbf{B} = \sum_{\ell=1}^{g} n_\ell (\bar{\mathbf{x}}_\ell - \bar{\mathbf{x}})(\bar{\mathbf{x}}_\ell - \bar{\mathbf{x}})' \text{ and } \mathbf{W} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_\ell)(\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_\ell)'$$

We have used

$$\text{Wilks lambda statistic } \Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|}$$

which is equivalent to the likelihood ratio test.

Other Multivariate Statistics

Three other multivariate test statistics are regularly included in the output of statistical packages

$$\text{Lawley-Hotelling Trace} = tr[\mathbf{B}\mathbf{W}^{-1}]$$
$$\text{Pillai trace} = tr[\mathbf{B}(\mathbf{B}+\mathbf{W})^{-1}]$$
$$\text{Roy's largest root} = \text{maximum eigenvalue of } \mathbf{W}(\mathbf{B}+\mathbf{W})^{-1}$$

Equivalent Test Statistics

- All four of these tests appear to be nearly equivalent for extremely large samples.
- For moderate sample sizes, all comparisons are based on what is necessarily a limited number of cases studied by simulation.
- From the simulations reported, the first three tests have similar power, while the last, Roy's test, behaves differently.
- Its power is best only when there is a single nonzero eigenvalue and, at the same time, the power is large.

- There is also some suggestion that Pillai's trace is slightly more robust against nonnormality.
- All four statistics apply in the two-way setting and in even more complicated MANOVA.
- When, and only, when the multivariate tests signals a difference, or departure from the null hypothesis, do we probe deeper.

Pillai's Statistic

```r
summary(rootstock.model)  # default output
```

```
#               Df Pillai approx F num Df den Df Pr(>F)
# ind.variable  5    1.3     4.07     20    168  2e-07 ***
# Residuals    42
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Hotelling-Lawley's Statistic

```
summary(rootstock.model, test = "H")
```

```
#              Df Hotelling-Lawley approx F num Df den Df
# ind.variable  5             2.92     5.48     20    150 2
# Residuals    42
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Roy's Statistics

```r
summary(rootstock.model, test = "R")
```

```
#              Df  Roy approx F num Df den Df Pr(>F)
# ind.variable  5 1.88     15.8      5     42  1e-08 ***
# Residuals    42
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Post Hoc Tests: 4 ANOVA's on the four tree measurements

```
summary(aov(dep.variable ~ ind.variable))[1:2]
```

```
#  Response Girth.4y :
#              Df Sum Sq Mean Sq F value Pr(>F)
# ind.variable  5  0.074 0.01471    1.93   0.11
# Residuals    42  0.320 0.00762
#
#  Response Growth.4y :
#              Df Sum Sq Mean Sq F value Pr(>F)
# ind.variable  5    4.2   0.840    2.91  0.024 *
# Residuals    42   12.1   0.289
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

```r
summary(aov(dep.variable ~ ind.variable))[3:4]
```

```
#  Response Girth.15y :
#              Df Sum Sq Mean Sq F value  Pr(>F)
# ind.variable  5   6.11   1.223      12 3.1e-07 ***
# Residuals    42   4.29   0.102
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
#
#  Response WgtAbvGrnd.15y :
#              Df Sum Sq Mean Sq F value  Pr(>F)
# ind.variable  5   2.49   0.499    12.2 2.6e-07 ***
# Residuals    42   1.72   0.041
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Except for trunk girth at four years, there are significant differences in the means of rootstock measurements amongst the six groups.